Acoustic Gait-based Person Identification using Hidden Markov Models

Jürgen T. Geiger, Maximilian Kneißl, Björn Schuller[1] *and Gerhard Rigoll*

Institute for Human-Machine Communication, Technische Universität München, Munich, Germany
[1]also with the Department of Computing, Imperial College London, London, U.K.
geiger@tum.de

Abstract

We present a system for identifying humans by their walking sounds. This problem is also known as acoustic gait recognition. The goal of the system is to analyse sounds emitted by walking persons (mostly the step sounds) and identify those persons. These sounds are characterised by the gait pattern and are influenced by the movements of the arms and legs, but also depend on the type of shoe. We extract cepstral features from the recorded audio signals and use hidden Markov models for dynamic classification. A cyclic model topology is employed to represent individual gait cycles. This topology allows to model and detect individual steps, leading to very promising identification rates. For experimental validation, we use the publicly available TUM GAID database, which is a large gait recognition database containing 3 050 recordings of 305 subjects in three variations. In the best setup, an identification rate of 65.5 % is achieved out of 155 subjects. This is a relative improvement of almost 30 % compared to our previous work, which used various audio features and support vector machines.

Index Terms: Acoustic gait-based person identification, gait recognition, hidden Markov models

1. Introduction

Recognising people by the way they walk (also known as gait recognition or gait-based person identification) is a relatively new field of research. Most of previously studied methods work in the visual domain, where this topic is an active field of research since the last decade [1]. However, acoustic information can also be used for gait recognition. Even though the focus on this modality has so far been significantly less, results are promising. While in the visual domain, identification systems can rely on analysing the silhouette [2], the task is much more difficult for systems working only with audio information. The relevant information which can be exploited by such systems consists not only of the sounds of the steps, but also adjacent sounds produced by the clothes of moving arms and legs. These sounds are influenced by the gait pattern of the walking person, making them suitable to be used for person identification. Furthermore, the sounds produced during walking are highly dependent on factors such as the floor type, type of shoes and clothes.

In a user study [3], the potential of humans to recognise others by their walking sounds was evaluated. After a training phase, twelve subjects were able to identify their co-workers by their walking sounds with an accuracy of 66 %. This result shows that sounds produced by walking persons convey characteristic information about the subject and can thus be used for person identification.

Potential applications of gait-based person identification using audio information are smart homes for ambient assisted living, indoor surveillance scenarios, or access control systems. Such an audio-based system can be used to enhance visual surveillance and facilitate multimodal approaches. As compared to video-based person identification, acoustic systems will also work in the darkness, require less expensive hardware and often lower sensor density and are less obtrusive. Acoustic gait-based person identification is also known as *acoustic gait recognition*.

1.1. Contribution

The contribution of this paper is a system for acoustic gait-based person identification that is based on hidden Markov models (HMMs). To our knowledge, this is the first time that HMMs are applied for this task. We use Mel-frequency cepstral coefficients (MFCCs) as audio features and HMMs with a cyclic topology for dynamic classification, in order to model the dynamics of gait patterns. With the cyclic topology, one pass through the model corresponds to a half gait cycle containing one step. Thus, the system is capable of detecting the individual steps in a recording and using them for person identification. Experiments are conducted using the TUM GAID corpus, which contains 3 050 recordings of 305 subjects in three walking variations in a realistic setup. The recognition system is trained with normal walking style recordings and evaluated on other recordings of normal walking style as well as variations including a backpack and shoe covers. Our experimental results show that the developed system is capable of achieving excellent recognition rates compared to previous work.

1.2. Related Work

The most-widespread approach for video-based gait recognition is the Gait Energy Image (GEI) [4], which is a simple silhouette-based approach. It can be combined with face recognition [5] or with depth information [6]. Furthermore, model-based approaches have been proposed for visual gait recognition [7]. Besides using video or audio information, other methods to identify walking persons include using acoustic Doppler sonar [8] or pressure sensors in the floor [9].

Using audio information for the task of gait-based person identification is a relatively new research field. In [10], footstep sounds were detected in a corpus of various environmental sounds. A system for person identification using footstep detection was introduced in [11]. The system was tested with a database of five persons. This work was extended in [12] by adding psychoacoustic features such as loudness, sharpness, fluctuation strength and roughness. Finally, in [13], dynamic time warping was used for classification and the database was extended to contain ten persons. The system achieves almost 100 % perfect classification rates (using ten persons). However, the task is simplified by reducing it to classification of pre-segmented footsteps. A similar task is addressed in the re-

cently published study by Altaf et al. [14]. There, a database of segmented footstep sounds from ten persons is used. Instead of extracting spectral features, the shape and properties of a footstep sound are examined in a temporal energy domain. As a result, an identification accuracy above 90 % is achieved by using a large number of footsteps during testing. When using only three consecutive footsteps, which is more comparable to our work, an accuracy of 45 % is obtained. Other studies on acoustic gait-based person identification were presented in [15, 16]. The weakness of all previous studies about acoustic gait-based person identification that are mentioned here is the fact that only small databases (mostly no more than ten subjects) that are overly prototypical have been employed. In addition, very often, classification is performed using pre-segmented footsteps. In our previous work [17], we investigated the potential of spectral, cepstral and energy-related audio features in combination with support vector machines (SVM) for acoustic gait-based person identification. This work was continued in [18], where a feature analysis method was used to select relevant audio features. In [19], we had also employed cyclic HMMs, for animal sound classification. The cyclic model topology proved to be efficient to model the repetitive structure of these sounds.

The remainder of this paper is structured as follows: In Section 2, we introduce the TUM GAID database which is used in the experiments. The employed system is described in Section 3, followed by the experimental setup and results in Section 4. Some concluding remarks are given in Section 5.

2. The TUM GAID Database

For our experiments, we use our freely available[1] TUM Gait from Audio, Image and Depth (GAID) database [17]. The motivation behind the TUM GAID database is to foster research in multimodal gait recognition. Therefore, data was recorded with an RGB-D sensor, as well as with a four-channel microphone array. Thus, a typical colour video stream, a depth stream and an audio stream are simultaneously available. The database contains recordings of 305 subjects walking perpendicular to the recording device in a 3.5 m wide hallway corridor with a solid floor. In each recorded sequence, the subject walks for roughly 4 m, typically performing between 1.5 and 2.5 gait cycles (each of them consisting of two steps). Most of the sequences have a length of approximately 2 – 3 s. Three variations are recorded for each subject: Normal walking ($\mathcal{N}$), walking with a backpack ($\mathcal{B}$), and walking with shoe covers ($\mathcal{S}$). For each subject, all recordings of the $\mathcal{N}$ condition were recorded directly after each other. This means that the same shoes and clothes are used, which corresponds more to a re-identification scenario. The backpack constitutes a significant variation in gait pattern and sound, and the shoe covers pose a considerable change in acoustic condition. Figure 1 shows screenshots of the three different walking conditions for one subject. For each subject, there are six recordings of the $\mathcal{N}$ setup, and two each of the $\mathcal{B}$ and $\mathcal{S}$ setups. This sums to a total number of 3 050 recordings. The metadata distribution of the database is well-balanced with a female proportion of 39 % and ages ranging from 18 to 55 years (average 24.8 years and standard deviation 6.3 years). More than half of the subjects are wearing sneakers while other commonly-used types of shoes are boots and loafers.

To allow for a proper scientific evaluation and to prevent overfitting on the test data, the database is divided into a *development set* and a *test set*. The two sets are person-disjunct and contain 150 and 155 subjects, respectively. Both for the development and for the test set, the first four $\mathcal{N}$ recordings of each subject are used for the enrollment process. The other two $\mathcal{N}$ recordings as well as the $\mathcal{B}$ and $\mathcal{S}$ recordings are used to perform the identification experiments. This means that models are learnt only using the $\mathcal{N}$ recordings, while the $\mathcal{B}$ and $\mathcal{S}$ conditions constitute previously unseen variations during the identification experiments and will therefore deteriorate the identification performance. The partition of the database is shown in Table 1.

[1] www.mmk.ei.tum.de/tumgaid

(a) Normal recording (b) Backpack recording (c) Shoe cover recording

Figure 1: Screenshots of three recordings in the TUM GAID database

Table 1: Partition of the TUM GAID database

	Development (150 subj.)	**Test** (155 subj.)
$\mathcal{N}1$ – $\mathcal{N}4$	Enrollment	Enrollment
$\mathcal{N}5$ – $\mathcal{N}6$	Identification	Identification
$\mathcal{B}1$ – $\mathcal{B}2$	Identification	Identification
$\mathcal{S}1$ – $\mathcal{S}2$	Identification	Identification

3. System Description

We use an HMM system for classification. Each individual subject is modelled by one HMM. While we started with using system settings from a simple word-based speech recognition system, we modified and improved the system properties to fit to the problem of acoustic gait recognition.

3.1. Audio Features

In our previous work we focussed on exploring the suitability of different audio features for the problem of acoustic gait-based person identification [18]. Using SVMs for classification, we evaluated different feature sets containing MFCCs and other spectral or energy-related features. Since SVMs are relatively robust (in contrast to HMMs) with regard to the number of employed features, we were able to improve the average identification accuracy (on the test set of the TUM GAID database) from 23.9 % (only MFCCs) to 28.2 % by adding and selecting relevant features. In the present work, the focus is not on the front-end processing but rather on the back-end recognition system. Therefore we keep the front-end fixed to using only MFCCs. We use MFCC features in the standard configuration: MFCCs 0–12 including their delta and acceleration coefficients, computed every 10 ms from a 25 ms Hamming window, resulting in 39 features in total. While the database provides four-channel audio recordings, we extract features from monaural recordings, which are obtained by averaging over the four channels. In addition, we obtained slight improvements by processing the audio features with principal component analysis (PCA), without reducing the number of components. Here, the transformations are computed only on the enrollment data, and applied on both

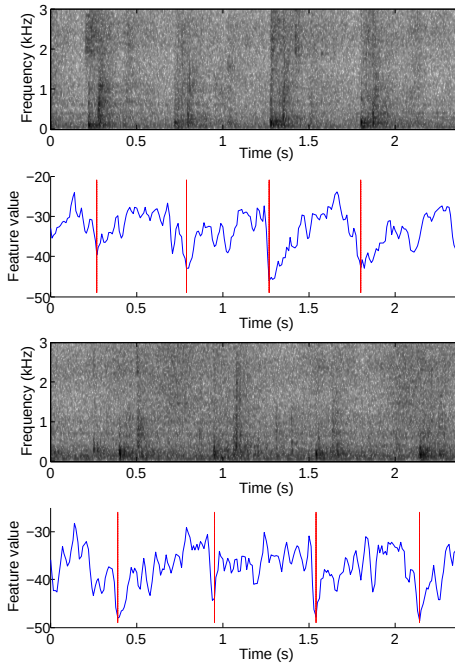

Figure 2: Spectrograms (top) and corresponding first MFCC coefficients (bottom), each, for a normal-type recording of two different subjects. Temporal position of footsteps is marked with a vertical line.

the enrollment and identification data.

Figure 2 shows the spectrograms and corresponding first MFCC coefficients for two exemplary recordings ($\mathcal{N}$ setup) of two different subjects. The spectrograms reveal a considerable static background noise, which is due to the recording environment. Several spectral peaks can be identified which correspond to the footsteps and the sounds between the steps, which are mostly made by the legs of the trousers or skirts rubbing against each other. In the plot of the MFCCs, the temporal position of the steps are marked. The behaviour of the MFCC features indicates that they are useful to detect the position of the steps and to distinguish between different persons.

3.2. HMM System

Our starting point is a simple HMM system that can be compared to a whole-word recognition system (each person representing one *word*) in speech recognition. Each subject in the dataset is represented by an HMM. The models are equipped with a linear left-right topology. With such a model topology, the HMM has to pass through all of its states sequentially without skipping a state. Before introducing an appropriate step modelling method, which will be described in the next subsection, we apply an approach where each recording containing several steps is modelled by one pass through an HMM. As a result, rather large numbers of states (generally more than ten) are required to be able to model the dynamic sequence of sounds during walking.

In a standard HMM system, the observations are modelled with a mixture of Gaussians. However, our first experiments showed that the best results are obtained by using HMMs with a single Gaussian state model, as the amount of training data is very small and hence probably not sufficient to train a more fine-grained distribution of the features. Another reason could be that a higher number of components leads to overfitting, modelling also the noise in the recordings.

During decoding, a grammar controls the possible recognition output. Our most simple employed grammar follows the basic HMM system setup where exactly one pass through a model is allowed for each recording. A multi-step grammar is then introduced to let the system automatically segment the recording: Any number of repetitions of the same model (subject) is allowed. In order to train the HMMs to model the separate steps, an approach using a cyclic HMM topology is employed as described in the following.

3.3. Step Modelling

To be able to model the individual steps in each recording, we use cyclic HMMs. In our basic HMM system, each recording (containing several gait cycles) is modelled by one pass through the HMM. The strategy of representing each gait cycle separately by one pass through all states of the HMM is better suited to model the observations. We consider the two halves of each gait cycle to be equivalent (although in fact, there is a person-dependent assymetry [20]), and therefore the system is designed to model half a gait cycle (containing one step) by each HMM. In this way, one pass through the HMM models the sounds of one step and adjacent sounds (produced by the moving arms and legs). This method of step modelling is implemented in the system configuraton and training in the following way: The state transition matrix of each HMM has a left-right topology, and jumps from the last state to the first state are allowed. Models are trained with embedded re-estimation, where the number of steps is known (as determined by simple video processing methods). As a result, the position of the steps in the training data is automatically estimated during model training. Together with the introduced multi-step decoding grammar, the developed system is then capable of detecting, segmenting and recognising the steps occurring in the recordings.

4. Experiments

Experiments are performed with the TUM GAID database that was described in Section 2, using the development set for system design and tuning. Finally, we use the test set to evaluate our best system configuration. For all systems evaluated using the development set, 15 HMM states appeared to be the optimal configuration. In addition, the best results were obtained with six training iterations. For each system setup, we report experimental results (identification accuracy) separately for the three different recording conditions (normal, backpack, shoe covers). In addition, the average accuracy over these three conditions is included.

4.1. Development set

Table 2 shows the results on the development set for different system configurations. The basic HMM system without explicit modelling of separate steps (cf. Section 3.2) is the first evaluated system. In the normal recording condition, slightly more than half of the testing samples are classified correctly. Averaging over the three different conditions, an accuracy of 28.2 % is obtained, which serves as a baseline for further experiments. The first step towards the improved recognition system is the introduction of a decoding grammar which allows to recognise multiple sequential instances of the same subject in the recordings. This modification improves the average accuracy to

Table 2: *Development set (150 subjects) evaluation of different audio features, for the normal ($\mathcal{N}$), backpack ($\mathcal{B}$) and shoe cover ($\mathcal{S}$) recording conditions.*

	Condition			
Accuracy [%]	$\mathcal{N}$	$\mathcal{B}$	$\mathcal{S}$	**average**
basic HMM	53.3	30.7	7.0	28.2
+ multi-step decoding	56.3	31.3	7.3	31.6
+ PCA	57.7	34.3	9.7	33.9
+ step modelling	69.7	44.7	9.3	41.2

Table 3: *Test set (155 subjects) evaluation of our system compared to our previously published results, for the normal ($\mathcal{N}$), backpack ($\mathcal{B}$) and shoe cover ($\mathcal{S}$) recording conditions.*

	Condition			
Accuracy [%]	$\mathcal{N}$	$\mathcal{B}$	$\mathcal{S}$	**average**
video (GEI) [17]	99.4	27.1	52.6	59.7
baseline SVM [17]	44.5	27.4	4.8	25.6
SVM + feat. sel. [18]	51.9	28.4	4.2	28.2
basic HMM	41.0	24.2	7.1	24.1
improved HMM	65.5	36.5	9.0	37.0

31.6 % (mostly due to improvements in the $\mathcal{N}$ setup). Applying PCA to the features improves the accuracy for all three recording conditions. Training the system to model each step by one pass through an HMM (cf. Section 3.3) leads to the largest improvement in accuracy. In the normal walking condition, more than two thirds of the samples are now identified correctly. The accuracy in the backpack walking condition is also greatly improved, whereas the performance in the shoe cover condition remains largely unaffected. While the improvements obtained with the multi-step grammar and PCA are not significant, improved step modelling leads to a significant improvement in the $\mathcal{N}$ and $\mathcal{B}$ conditions and for the average accuracy (evaluated with a one-tailed t-test with a significance level of $\alpha = 0.05$).

With a simple analysis we examined the system's ability to correctly detect the individual steps. To this end, we use the best-performing developed system (row four in Table 2). For the test samples of the normal walking conditions, we observe the number of steps detected by the system. The average number of steps in these test recordings is 5.3, while the system predicts 4.3 steps, on average. For correctly identified *subjects*, the average number of predicted *steps* is 5.0, while for incorrectly identified subjects it is 3.5. This shows that when the subjects are identified correctly, the step segmentation works very well.

4.2. Test set

In Table 3, we show the results on the test set, for our baseline system and the best system configuration. For comparison, we include our previously published results on the same dataset. The first row shows results of a state-of-the-art gait recognition method working with video data, namely the GEI [17]. This method achieves almost perfect results in the normal walking condition, while especially the backpack and also the shoe variation constitute a real difficulty for the system (59.7 % on average). However, these results have to be interpreted carefully, since the GEI utilises mainly the appearance (the silhouette of a person) and not the behaviour (the gait pattern). Using a large set of different audio features (1 625 static features per recording) and SVMs for classification (second row) was our first audio-domain baseline system [17]. Naturally, the addressed task is much more difficult when dealing only with audio data (average accuracy 25.6 %). However, this system can compete with the GEI in the backpack recording variation. In [18], we improved the SVM system by employing a feature-selection technique to chose relevant features for the task, obtaining an average identification accuracy of 28.2 %. Now, with our basic HMM setup, the resulting accuracy of 24.1 % is comparable to the baseline SVM system. The methods introduced in this work (primarily modelling each step separately during model training and decoding) are able to bring a large improvement, reaching 37.0 %. In the $\mathcal{N}$ and $\mathcal{B}$ recording conditions, the accuracy is improved significantly, by more than one third. The accuracy of the video-processing method (GEI) in the backpack recording condition is surpassed by 26 % relatively. Compared to the previous best-performing audio system (the SVM system including feature selection) the average accuracy is improved by 24 %, relatively (significant in all recording conditions).

5. Conclusions

We developed a model-based system for recognising people from walking sounds. The system uses HMMs in a cyclic topology to automatically segment the recordings according to separate steps. Experiments were conducted using the TUM GAID database containing recordings of 305 subjects (150 in the development set and 155 in the test set) in three different recording conditions: normal walking, walking with a backpack, and walking with shoe covers. The results show that a basic HMM system (without explicit modelling of separate steps) achieves a similar performance in comparison to the SVM system presented in our previous work. Improving the system with the methods introduced in this work results in large performance gains in identification accuracy. With this system, each half gait cycle is modelled by one pass through a cyclic HMM. This covers the sound of one step and adjacent sounds, which are mainly produced by moving arms and legs. Thus, it is clear that the backpack or shoe cover variation influence the identification performance in a negative way. However, when identification experiments are carried out with the same walking style and shoe type as the model was trained with (normal walking condition), almost two thirds of the subjects are identified correctly from the test set containing 155 individuals.

Given the challenging but application-friendly enrollment of only four examples per walking subject and in order to improve the robustness of the system, adopting approaches from speaker recognition like creation of models through adaption from a background model [21] could be a promising strategy in the future. Furthermore, we will work on improving the system's robustness to variations. This includes better coping with the backpack and shoe cover recording conditions. In addition, the TUM GAID database contains a set of subjects with recordings made on two different dates in time (with three months in between). Therewith, the influence of changing types of shoes and clothes as well as possibly higher variation of the walking style on the system performance can be evaluated. In order to improve the system in this direction, we want to test approaches to address session variability known from speaker recognition (such as joint factor analysis [22]) as well as methods for model adaptation or feature transformation adopted from speech recognition systems.

6. References

[1] L. Lee and W. Grimson, "Gait analysis for recognition and classification," in *Proc. IEEE International Conference on Automatic Face and Gesture Recognition*, Washington, D.C., USA, 2002, pp. 148–155.

[2] L. Wang, T. Tan, H. Ning, and W. Hu, "Silhouette analysis-based gait recognition for human identification," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 25, no. 12, pp. 1505–1518, 2003.

[3] K. Mäkelä, J. Hakulinen, and M. Turunen, "The use of walking sounds in supporting awareness," in *Proc. International Conference on Auditory Display*, Boston, MA, USA, 2003, pp. 144–147.

[4] J. Han and B. Bhanu, "Individual recognition using gait energy image," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 28, no. 2, pp. 316–322, 2006.

[5] M. Hofmann, S. Schmidt, A. Rajagopalan, and G. Rigoll, "Combined face and gait recognition using alpha matte preprocessing," in *Proc. IAPR/IEEE International Conference on Biometrics*, New Delhi, India, 2012, pp. 1–6.

[6] M. Hofmann, S. Bachmann, and G. Rigoll, "2.5D gait biometrics using the depth gradient histogram energy image," in *Proc. IEEE International Conference on Biometrics: Theory, Applications and Systems*, Washington, DC, USA, 2012.

[7] C. Yam, M. Nixon, and J. Carter, "Automated person recognition by walking and running via model-based approaches," *Pattern Recognition (Elsevier)*, vol. 37, no. 5, pp. 1057–1072, 2004.

[8] K. Kalgaonkar and B. Raj, "Acoustic doppler sonar for gait recogination," in *Proc. IEEE Conference on Advanced Video and Signal Based Surveillance*, London, UK, 2007, pp. 27–32.

[9] J. Yun, S. Lee, W. Woo, and J. Ryu, "The user identification system using walking pattern over the ubifloor," in *Proc. International Conference on Control, Automation, and Systems*, Gyeongju, Korea, 2003, pp. 1046–1050.

[10] B. She, "Framework of footstep detection in in-door environment," in *Proc. International Congress on Acoustics*, Kyoto, Japan, 2004, pp. 715–718.

[11] Y. Shoji, T. Takasuka, and H. Yasukawa, "Personal identification using footstep detection," in *Proc. IEEE International Symposium on Intelligent Signal Processing and Communication Systems*, Seoul, South Korea, 2004, pp. 43–47.

[12] A. Itai and H. Yasukawa, "Footstep recognition with psyco-acoustics parameter," in *Proc. IEEE Asia Pacific Conference on Circuits and Systems*, Singapore, 2006, pp. 992–995.

[13] ——, "Footstep classification using simple speech recognition technique," in *Proc. IEEE International Symposium on Circuits and Systems*, Seattle, WA, USA, 2008, pp. 3234–3237.

[14] M. Altaf, T. Butko, and B.-H. Juang, "Person identification using biometric markers from footstep sounds," in *Proc. Interspeech*, Lyon, France, 2013, pp. 2934–2938.

[15] R. de Carvalho and P. Rosa, "Identification system for smart homes using footstep sounds," in *Proc. of IEEE International Symposium on Industrial Electronics*, Bari, Italy, 2010, pp. 1639–1644.

[16] D. Alpert and M. Allen, "Acoustic gait recognition on a staircase," in *Proc. IEEE World Automation Congress*, 2010, pp. 1–6.

[17] M. Hofmann, J. Geiger, S. Bachmann, B. Schuller, and G. Rigoll, "The TUM Gait from Audio, Image and Depth (GAID) Database: Multimodal Recognition of Subjects and Traits," *Journal of Visual Communication and Image Representation (JVCI), Special Issue on Visual Understanding and Applications with RGB-D Cameras*, vol. 25, no. 1, pp. 195–206, 2014.

[18] J. T. Geiger, M. Hofmann, B. Schuller, and G. Rigoll, "Gait-based person identification by spectral, cepstral and energy-related audio features," in *Proc. ICASSP*, Vancouver, Canada, 2013, pp. 458–462.

[19] F. Weninger and B. Schuller, "Audio recognition in the wild: Static and dynamic classification on a real-world database of animal vocalizations," in *Proc. ICASSP*, Prague, Czech Republic, 2011, pp. 337–340.

[20] M. Nixon, T. Tan, and R. Chellappa, *Human identification based on gait*. Springer, 2006, vol. 4.

[21] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital signal processing*, vol. 10, no. 1, pp. 19–41, 2000.

[22] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 4, pp. 1435–1447, 2007.